View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Copenhagen Business School: CBS Open Journals

LSP Professional Communication Knowledge Management Cognition
LSP Journal, Vol.2, No.2 (2011) / http://lsp.cbs.dk

CAOS – Construction and validation of Terminological Ontologies

Hanne Erdman Thomsen and Bodil Nistrup Madsen
Department of International Language Studies and Computational Linguistics
Copenhagen Business School, Denmark
het.isv@cbs.dk / bnm.danterm@cbs.dk

Abstract

This paper presents some principles of terminological ontologies implemented in the prototype that has been developed in the research project CAOS - Computer-Aided Ontology Structuring. Furthermore, some issues that have to be faced to further develop facilities for automatic consistency checking and automatic changes to ontologies, are discussed. The presentation will illustrate central facilities of the current version of the CAOS prototype, which is interactive and presupposes an end-user with a background in terminology rather than in formal ontology.

Introduction

A terminological ontology is a domain specific ontology, cf. for example the categorization of ontologies by Guarino (1998). We use the term *terminological ontology* as synonym to the term *concept system*, which is normally used in terminology work, cf. for example (ISO 704, 2000).

The principles of terminological ontologies presented here, build on the principles of terminology work as presented in (ISO 704, 2000), but have been further developed in the research and development project CAOS - Computer-Aided Ontology Structuring - whose aim is to develop a computer system designed to enable semi-automatic construction of concept systems, or ontologies, cf. (Madsen et al., 2005).

Terminological ontologies model concepts and the relations between them, and a concept is described by means of characteristics that denote properties of individual referents belonging to the extension of that concept. Other ontologies most commonly model classes, described by means of properties, and the relations between classes.

It is possible to use all types of concept relations in CAOS. The system offers a set of concept relations organized in a taxonomy, cf. (Madsen et al., 2002). It is also possible for the user to introduce user defined relations. For other presentations of concept relations, see for example (Nuopponen, 2005).

The CAOS Prototype

The backbone of terminological concept modeling in CAOS is constituted by characteristics modeled by formal feature specifications, i.e. attribute-value pairs, cf. (Carpenter, 1992). The use of feature specifications is subject to a number of principles and constraints.

Figure 1 presents part of an ontology for prevention created in CAOS. As can be seen, the graphical presentation is UML-based.

Consistency checking in CAOS

The technology developed in CAOS enables validation of the inheritance of characteristics when a new concept is introduced into a concept system. In a type hierarchy, subordinate concepts inherit characteristics from their superordinate concepts, and hence it is possible to validate whether the position of a given concept allows for the characteristics associated with it.

The facilities for semi-automatic construction of ontologies and for consistency checking in CAOS are, among other things, based on the introduction of dimensions and dimension specifications. A dimension of a concept is an attribute occurring in a (non-inherited) feature specification of one of its subordinate concepts, i.e. an attribute whose possible values allow a distinction between some of the subconcepts of the concept in question. A dimension specification consists of a dimension and the values associated with the corresponding attribute in the feature specifications of the subordinate concepts: dimension: [value1| value2| ...]. In this way, the principle of subdivision criteria that has been used for many years in terminology work, has been formalized in CAOS.

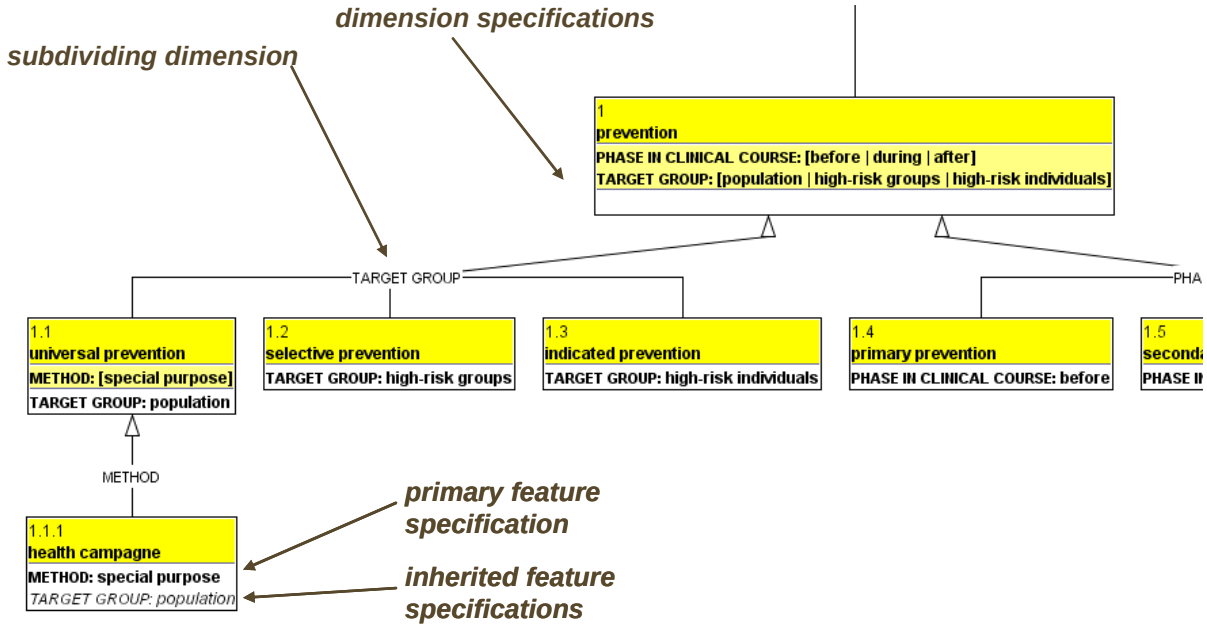

Figure 1. Extract of an ontology for prevention

One or more of the dimensions of a concept must be chosen as the subdividing dimensions. Subdividing dimensions must be chosen in such a way that each daughter concept has one and only one feature specification containing as an attribute a subdividing dimension of the

mother concept. This ensures that there are no overlapping subdividing dimensions, and hence no overlap in partitions.

In the following, a brief description of some important principles of CAOS will be given:

- grouping by subdividing dimensions, including choice of subdividing dimensions and no overlapping of subdividing dimensions,
- uniqueness of primary feature specifications and
- uniqueness of dimensions.

Grouping by subdividing dimensions

From figure 1 it is seen that *prevention* may differ with respect to both target group and phase in clinical course. However, in the case of the three concepts *universal prevention*, *selective prevention* and *indicated prevention* it is obvious that TARGET GROUP must be chosen as the subdividing dimension (subdivision criterion). If the user tries to choose a second dimension as subdividing dimension for the three mentioned subordinate concepts, CAOS will not allow it, and will consequently warn the user. The feature specifications comprising the subdividing dimension (referred to as the delimiting feature specifications) will form the basis for the definition of the three concepts.

Constraints in CAOS related to subdivision criteria are:

- A concept (with only one mother concept) may contain at most one delimiting feature specification
- A concept (of level 2 or below) must contain at least one delimiting feature specification

Another constraint is that an attribute may only be associated with one value in a feature structure on a given concept (a combination of two or more feature specifications on a concept is called a feature structure). If the user attempts to create a concept *universal selective prevention* with two superordinate concepts within the same group (dimension: TARGET GROUP), this would mean that the attribute TARGET GROUP would be associated with two values in the feature structure for *universal selective prevention*: TARGET GROUP: population and TARGET GROUP: high-risk groups. CAOS will not allow this 'illegal polyhierarchy'. This type of error is also known as a partition error (Góméz-Pérez et al. 2003).

In Protégé[1] this can be handled be adding a new superordinate concept to a concept on the basis of the formal definition of the concepts in question. However, this treatment is not feasible for the end users we have in mind, who have no training in formal logic or similar.

Uniqueness of dimensions

The principle of uniqueness of dimensions states that a given dimension may occur on only one concept in an ontology. Uniqueness of dimensions helps to create coherence and simplicity in the ontological structure because concepts that are characterised by means of primary feature specifications with the same dimension must appear as coordinate concepts on the same level having a common superordinate concept.

[1] http://protege.stanford.edu/

Uniqueness of feature specifications
The principle of uniqueness of feature specifications stipulates that a feature specification may occur only once in a terminological ontology as primary. A primary feature specification is entered on a concept directly by the terminologist, as opposed to inherited feature specifications, which are inherited from superordinate concepts.

Uniqueness of dimensions (the previous principle) means that a given primary feature specification can only appear on concepts that are daughters of the concept containing the relevant dimension. Uniqueness of primary feature specifications means that a given primary feature specification can only appear on one of these daughters. If the terminologist tries to insert the primary feature specification [TARGET GROUP: population] on the concept *selective prevention*, CAOS will report that [TARGET GROUP: population] is already specified on the concept 1.1 *universal prevention*.

The motivation of the principle of uniqueness of primary feature specifications is that

- characteristics will always serve to distinguish concepts, and
- common characteristics should be located on a common superordinate concept (this principle may contribute to the identification of potential gaps in the ontology).

Characteristics of the CAOS prototype compared to other ontology editors
Several other tools for creating ontologies have been (or are being) developed, e.g. Protégé and WebODE[2].

The main difference between the system for terminological ontologies, described here, and other systems is that in the latter, terminological information cannot be modeled and presented in the same way. This information, i.e. subdivision criteria and dimension specifications, is crucial in the development of terminological ontologies. Furthermore, in order to check conformance to the constraints mentioned in section 2.2 – 2.4, the end user must be able to formulate formal constraints for each subdivision criterion. In CAOS, the constraints are part of the system.

Further Development of the CAOS Prototype

In a new project we aim to develop an additional prototype that will be able to automatically build a first draft ontology on the basis of a domain-specific text corpus. This prototype will be based on a combination of existing and new methods and principles for automatic extraction of concepts and information about concepts, i.e. characteristics and concept relations.

Another aim is to further develop CAOS so that it may be used for automatic validation of draft ontologies that are the result of the automatic knowledge extraction described above. The new prototype will not just be able to detect errors, it will also propose corrections of errors. For example it will automatically handle partition errors. To our knowledge no other systems have such capabilities.

To further develop facilities in CAOS for automatic consistency checking and automatic changes to ontologies, various issues have to be dealt with.

[2] http://mayor2.dia.fi.upm.es/oeg-upm/index.php/en/downloads/60-webode

Validation of an ontology vs. validation of one concept
First of all, the technology currently used in CAOS validates one concept at a time, while the new prototype will need to validate an entire ontology provided by the knowledge extraction module.

Characteristics vs. relations
In CAOS, a concept may have both feature specifications and relations to other concepts. However, a given characteristic of a concept can be modeled either as an attribute-value pair or a relation-concept pair, e.g. in Figure 1, the characteristic modeled by the feature specification [TARGET GROUP: population] could have been modeled as a relation (HAS_TARGET_GROUP) to another concept (*population*).

The ontology extraction module will not be able to distinguish between attributes and relations. Therefore, in the new prototype, relations (other than type relations) and attributes of characteristics will have to be treated identically. In the validation they will be treated as attributes of characteristics, and the related concepts will be treated as values. This raises a theoretical research issue: is it necessary to differentiate relations and characteristics? If so, what is the difference?

Multiple values
A problem related to the above is that the CAOS technology allows a given concept to have only one value for a given attribute, while it may be related to several other concepts with the same relation. The extraction tool is bound to deliver more than one concept for a given relation (or value for a given attribute) for any concept. The CAOS technology needs to be modified to handle this.

Some relations may only be applied to a given concept once. For example, no concept can have more than one instance of the relation HAS_LENGTH_IN_CM. This corresponds to the CAOS principle mentioned above, i.e. that for a given attribute a concept can have at most one value. Hence a research issue to be investigated is whether these relations can be distinguished from those allowing for multiple instances, since this is important for validation.

Specialized values
An issue relating to characteristics is that of specialized values. In order to handle this, the CAOS technology needs to be enhanced to include a type hierarchy of values (or related concepts). The use of value hierarchies has been implemented e.g. in the Lexical Knowledge Base system (LKB) first developed by Ann Copestake for lexical semantics and further enhanced for HPSG[3] purposes, c.f. (Copestake, 1993).

Automatic positioning
A prerequisite for making automatic changes in the ontology based on the validation is to be able to position a concept in an existing type hierarchy by employing the characteristics registered for that concept. Techniques for positioning concepts and making automatic changes to the ontology are to be developed.

[3] Head Driven Phrase Structure Grammar

Perspectives

Terminological ontologies offer very detailed information about concepts, e.g. feature specifications, subdivision criteria and dimension specifications. The question is whether this information is useful in the various applications of ontologies. Undoubtedly, this information is needed for concept clarification, for example with a view to the definition of central concepts in the use of IT systems for information storage and retrieval.

In the SIABO project, Semantic Information Access through Biomedical Ontologies, cf. http://siabo.org, it is planned to test whether terminological ontologies will also add value to systems for ontology-based information retrieval.

References

Carpenter, Bob. 1992. *The Logic of Typed Feature Structures.* Cambridge University Press, UK.

Copestake, Ann. 1993 *The Compleat LKB, Technical Report No. 316*, University of Cambridge.

CWA 15045. 2005. *CEN Workshop Agreement: Multilingual Catalogue Strategies for eCommerce and eBusiness*.

Gómez Pérez, Asunción, Mariano Fernández-López, Oscar Corcho: (2003) "Ontological Engineering". Advanced Information and Knowledge Processing series. ISBN 1-85233-551-3. Springer Verlag.

Guarino, Nicola. 1998. Formal Ontology and Information Systems. *Formal Ontology in Information Systems, Proceedings of the First International Conference (FOIS'98)*, June 6-8, Trento, Italy, 3-15. Ed. Nicola Guarino. Amsterdam: IOS Press.

ISO 704. 2000. *Terminology work — Principles and methods.* Genève: ISO.

Madsen, Bodil Nistrup, Hanne Erdman Thomsen & Carl Vikner. 2005. Multidimensionality in terminological concept modelling. Bodil Nistrup Madsen, Hanne Erdman Thomsen (eds.): *Terminology and Content Development, TKE 2005, 7th International Conference on Terminology and Knowledge Engineering*, Copenhagen: 161-173.

Madsen, Bodil Nistrup, Bolette Sandford Pedersen & Hanne Erdman Thomsen. 2002. The Role of Semantic Relations in a Content-based Querying System: a Research Presentation from the OntoQuery Project. Simov, Kiril & Atanas Kiryakov (eds.): *Proceedings from OntoLex '2000, Workshop on Ontologies and Lexical Knowledge Bases,* Sept. 8-10 2000, Sozopol, Bulgaria: 72-81.

Nuopponen, Anita. 2005. Concept Relations. Bodil Nistrup Madsen, Hanne Erdman Thomsen (eds.): *Terminology and Content Development, TKE 2005, 7th International Conference on Terminology and Knowledge Engineering*, Copenhagen, 127-138.

***